Law of Large Numbers and the Central Limit Theorem in Number Theory

Anish Ray
University of Münster

20 December 2022

Introduction

In this talk we will explore the analogues of the WLLNs and CLT in number theory. We recall the weak law of large numbers which states that:

Theorem. ***(Khinchin's WLLNs)*** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of integrable, real-valued random variables that are pairwise uncorrelated. If*

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i)=0,$$

then

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\mathbb{E}\sum_{i=1}^{n}X_i\right|>\varepsilon\right)=0.$$

and the central limit theorem states that:

Theorem. ***(Lindeberg's CLT)*** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables such that $\forall n\in\mathbb{N}$, $0<\mathbb{V}X_n<\infty$ and $\sigma_n=\sqrt{\sum_{i=1}^{n}\mathbb{V}X_i}$. Then $\forall a,b\in\mathbb{R}$,*

$$\lim_{n\to\infty}\mathbb{P}\left(a\leq\frac{\sum_{i=1}^{n}(X_i-\mathbb{E}X_i)}{\sigma_n}\leq b\right)=\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{y^2}{2}}\,\mathrm{d}y,$$

if the Lindeberg's condition with $\eta_i=\mathbb{E}X_i$, $i=1,2,\ldots,$ as follows

$$\lim_{n\to\infty}\frac{1}{\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i-\eta_i)^2\mathbb{I}_{\{|X_i-\eta_i|\geq\varepsilon\sigma_n\}}\right]=0$$

is satisfied $\forall\varepsilon>0$.

Next, we need some definitions.

Definition 0.1. *A number theoretic function or arithmetic or arithmetical function f has domain $\mathbb{N}$ and its range is a subset of $\mathbb{C}$. Further, f is additive if $f(mn)=f(m)+f(n)$ whenever $\gcd(m,n)=1$.*

Definition 0.2. *The mean of f denoted by $M\{f(n)\}$ is defined as the limit (if it exists)*

$$M\{f(n)\} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(n).$$

Definition 0.3. *If A is a set of positive integers, we denote by $A(N)$ the number of its elements among the first N integers. If*

$$\lim_{N\to\infty} \frac{A(N)}{N} = D\{A\}$$

exists, it is called the density of A.

Consider the integers divisible by a prime p. The density of the set of these integers is clearly $\frac{1}{p}$. To be divisible by prime numbers p and q the integer should be divisible by pq, and consequently the density of the new set is $\frac{1}{pq}$. Since $\frac{1}{pq} = \frac{1}{p}.\frac{1}{q}$, one can interpret that the **"events"** of being divisible by two distinct prime numbers are independent. Of course, this holds for any number of prime numbers. This simple observation leads us to the foundational theory of Probabilistic Number Theory which we will discuss shortly in the next section.

Definition 0.4. *For a prime number p, the function $\rho_p(n)$ be defined as follows:*

$$\rho_p(n) = \begin{cases} 1 & \text{if} \quad p \mid n \\ 0 & \text{if} \quad p \nmid n \end{cases}, \forall n \in \mathbb{N}.$$

Observe now that if $\varepsilon_j = 0$ or 1, then for prime numbers $p_1, p_2, \ldots, p_k$, $D\{\rho_{p_1}(n) = \varepsilon_1, \rho_{p_2}(n) = \varepsilon_2, \ldots, \rho_{p_k}(n) = \varepsilon_k\} = D\{\rho_{p_1}(n) = \varepsilon_1\} D\{\rho_{p_2}(n) = \varepsilon_2\} \ldots D\{\rho_{p_k}(n) = \varepsilon_k\}$. This is simply another way of stating that the **"events"** of being divisible by $p_1, p_2, \ldots, p_k$ are independent or that the functions $\rho_p(n)$ are independent.

Definition 0.5. *An additive number-theoretic function f is called strongly additive if $f(p^\alpha) = f(p)$, where p is a prime and $\alpha = 2, 3, \ldots$. Further, $f(n) = \sum_p f(p)\rho_p(n)$.*

§1 The prime divisor counting function and the analogue of WLLNs

Definition 1.1. *Let $\omega(n)$ denote the number of prime divisors of n counting multiplicity, i.e, if $n = \prod_p p^{\alpha_p(n)}$, then*

$$\omega(n) = \sum_p \alpha_p(n).$$

Further, let $\nu(n)$ denote the number of prime divisors without counting multiplicity, i.e.,

$$\nu(n) = \sum_p \rho_p(n).$$

One can observe that $\nu(n)$ is a strongly-additive function.

As observed in the previous section, $\rho_p(n)$ are statistically independent and one might expect that the theory of addition of independent random variables can be applied to distribution problems of additive number-theoretic functions.
Since the density is only a finitely additive measure, probability theorems will be directly applicable only if a finite number of ρ_p's are involved. Thus one can apply probability theorems to **"truncated"** functions

$$f_k(n) = \sum_{p<k} f(p)\rho_p(n).$$

The difference $\omega(n) - \nu(n)$ will be called the excess, and we shall determine the density of integers for which the excess is equal to $k \in \mathbb{N} \cup \{0\}$, i.e.,

$$d_k = D\{\omega(n) - \nu(n) = k\}.$$

Since the existence of this density is not obvious and we need to be establish it. We start with the well-known formula, for $m \in \mathbb{Z}$

$$\frac{1}{2\pi}\int_0^{2\pi} e^{imx}\,\mathrm{d}x = \begin{cases} 1, & m = 0, \\ 0, & m \neq 0, \end{cases} \tag{1}$$

and consider

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\pi}\int_0^{2\pi} e^{i(\omega(n)-\nu(n)-k)x}\,\mathrm{d}x = \frac{1}{2\pi}\int_0^{2\pi} e^{-ikx}\frac{1}{N}\sum_{i=1}^{N} e^{i((\omega(n)-\nu(n))x}\,\mathrm{d}x. \tag{2}$$

The left hand side of (2) represents the fraction of integers $n \leq N$ whose excess is exactly k. Thus

$$d_k = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\pi}\int_0^{2\pi} e^{i(\omega(n)-\nu(n)-k)x}\,\mathrm{d}x \tag{3}$$

if the limit exists. Then from the bounded convergence theorem, it follows that it is enough to prove that for every $x \in \mathbb{R}$,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} e^{i((\omega(n)-\nu(n))x} = M\{e^{i(\omega(n)-\nu(n))x}\}. \tag{4}$$

Now $\omega(n) - \nu(n) = \sum\limits_p(\alpha_p(n) - \rho_p(n))$, and the functions $\alpha_p(n) - \rho_p(n)$ are easily seen to be independent. Then it follows, that $M\{e^{i(\omega(n)-\nu(n))x}\} = \prod\limits_p\left(1-\frac{1}{p}\right)\left(1+\frac{1}{p-e^{ix}}\right)$. I will not discuss the proof of the existence of the limit in (4) because it is rather long and uninteresting. Still, if anybody is interested then I ask you to refer to article [1] in the references. Anyway, eventually we obtain,

$$d_k = \int_0^{2\pi} e^{ikx}\prod_p\left(1-\frac{1}{p}\right)\left(1+\frac{1}{p-e^{ix}}\right)\mathrm{d}x.$$

Consider now the function

$$F(z) = \prod_p\left(1-\frac{1}{p}\right)\left(1+\frac{1}{p-z}\right),$$

and note that it is analytic in the whole plane, except for simple poles at the primes $z = 2, 3, 5, \ldots$. In particular, $F(z)$ is analytic in the circle $|z| < 2$, and we can expand F in a power series

$$F(z) = \sum_{k=0}^{\infty} a_k z^k$$

whose radius of convergence is 2. The coefficients a_k are given by

$$a_k = \frac{1}{2\pi i} \int \frac{F(z)}{z^{k+1}} \, \mathrm{d}z,$$

where the integral is taken over the circle $|z| = 1$, we obtain by substituting $z = e^{ix}$ that $a_k = d_k$. In other terms,

$$\sum_{k=0}^{\infty} d_k z^k = \prod_p \left(1 - \frac{1}{p}\right)\left(1 + \frac{1}{p - z}\right).$$

Setting $z = 1$, we obtain

$$\sum_{k=0}^{\infty} d_k = \prod_p \left(1 - \frac{1}{p}\right)\left(1 + \frac{1}{p - 1}\right) = 1.$$

Thus, d_k is analogous to a probability mass function of a discrete random variable.

The Hardy-Ramanujan Theorem

In 1917, Hardy and Ramanujan proved that almost every integer m has approximately $\log\log m$ divisors. The precise formulation is as follows:

Theorem 1.1. ***(Hardy-Ramanujan)*** *If a sequence $g_m \to \infty$ as $m \to \infty$ the density of integers for which*

$$\nu(m) < \log\log m - g_m\sqrt{\log\log m}$$

or

$$\nu(m) > \log\log m + g_m\sqrt{\log\log m}$$

is 0. Due to the slowness with which $\log\log m$ increases we have the equivalent formulation: If K_n is the number of integers m, $1 \leq m \leq n$, for which

$$\nu(m) < \log\log n - g_n\sqrt{\log\log n} \tag{5}$$

or

$$\nu(m) > \log\log n + g_n\sqrt{\log\log n}$$

then $\frac{K_n}{n} \to 0$ as $n \to \infty$.

The proof we are about to discuss is due to **Pál Turán**, and it is much simpler than the original proof of **Hardy** and **Ramanujan**. We will see that the proof resembles the idea of the application of Chebyshev's inequality in the WLLNs and is a direct analogue of that proof.

Proof. We have

$$\sum_{m=1}^{n} (\nu(m) - \log\log n)^2 \geq \sum_{m=1}^{n}{}' (\nu(m) - \log\log n)^2, \tag{6}$$

where the prime on the summation sign indicates that the summation is extended only over integers m satisfying (5). Clearly,

$$\frac{K_n}{n} \leq \frac{1}{ng_n^2 \log\log n} \sum_{m=1}^{n} (\nu(m) - \log\log n)^2. \tag{7}$$

Further, we obtain, that

$$\sum_{m=1}^{n} (\nu(m) - \log\log n)^2 = \sum_{m=1}^{n} \nu_m^2 - 2\log\log n \sum_{m=1}^{n} \nu(m) + n(\log\log n)^2, \tag{8}$$

where $\nu^2(m) = \sum\limits_{p} \rho_p(m) + 2\sum_{p<q} \rho_p(m)\rho_q(m)$ and consequently,

$$\sum_{m=1}^{n} \nu(m) = \sum_{p} \left\lfloor \frac{n}{p} \right\rfloor, \tag{9}$$

and

$$\sum_{m=1}^{n} \nu^2(m) = \sum_{p} \left\lfloor \frac{n}{p} \right\rfloor + 2\sum_{p<q} \left\lfloor \frac{n}{pq} \right\rfloor. \tag{10}$$

In (9) and (10) the summation is only over primes $p, q \leq n$, and thus

$$\sum_{m=1}^{n} \nu^2(m) \geq \sum_{p} \frac{1}{p} - \pi(n), \tag{11}$$

where $\pi(n) =$ no. of primes $p \leq n$. Similarly, we get

$$\sum_{m=1}^{n} \nu^2(m) \leq n \sum_{p \leq n} \frac{1}{p} + 2n \sum_{p<q \leq n} \frac{1}{pq} < n \sum_{p \leq n} \frac{1}{p} + n \left(\sum_{p \leq n} \frac{1}{p} \right)^2. \tag{12}$$

A result of **Merten's** gives us $\sum\limits_{p \leq n} \frac{1}{p} = \log\log n + e_n$ where e_n is bounded, and hence

$$\sum_{m=1}^{n} \nu^2(m) \leq n(\log\log n)^2 + 2ne_n \log\log n + ne_n^2 + n\log\log n + ne_n \tag{13}$$

and

$$\sum_{m=1}^{n} \nu(m) \geq n\log\log n + ne_n - \pi(n). \tag{14}$$

Finally, (8) yields

$$\sum_{m=1}^{n} (\nu(m) - \log\log n)^2 \leq ne_n^2 + n\log\log n + ne_n + 2\pi(n)\log\log n, \tag{15}$$

and consequently

$$\frac{K_n}{n} \leq \frac{1}{g_n^2} + \frac{e_n^2}{g_n^2 \log\log n} + \frac{e_n}{g_n^2 \log\log n} + 2\frac{\pi(n)}{n}\frac{1}{g_n^2}. \tag{16}$$

Since $\pi(n) < n$, the desired result follows as $n \to \infty$. □

§2 The Erdős-Kac Theorem a.k.a. the analogue of the CLT

The **Erdős-Kac** theorem given in 1939 by **Paul Erdős** and **Mark Kac,** is also known as the "fundamental theorem of probabilistic number theory", which can be loosely stated as $\mathbb{P}\left(n \leq N : \omega_1 \leq \frac{\nu(n)-\log\log n}{\sqrt{\log\log n}} \leq \omega_2\right) \to \frac{1}{\sqrt{2\pi}} \int_{\omega_1}^{\omega_2} e^{-\frac{y^2}{2}}\,\mathrm{d}y$ as $n \to \infty$ for all $\omega_1, \omega_2 \in \mathbb{R}$. This is an extension of the **Hardy-Ramanujan theorem** and marks the beginning of probabilistic number theory. Without further ado, let us state the precise theorem:

Theorem 2.1. *(Erdős-Kac) Let $K_n(\omega_1, \omega_2)$ be the number of integers m, $1 \leq m \leq n$, for which*

$$\log\log n + \omega_1\sqrt{\log\log n} \leq \nu(m) \leq \log\log n + \omega_2\sqrt{\log\log n}, \tag{17}$$

then

$$\lim_{n\to\infty} \frac{K_n(\omega_1,\omega_2)}{n} = \frac{1}{\sqrt{2\pi}} \int_{\omega_1}^{\omega_2} e^{-\frac{y^2}{2}}\,\mathrm{d}y, \forall \omega_1, \omega_2 \in \mathbb{R}. \tag{18}$$

Because of the slowness with which $\log\log n$ changes the result in (18) is equivalent to the statement:

$$D\{\log\log n + \omega_1\sqrt{\log\log n} \leq \nu(n) \leq \log\log n + \omega_2\sqrt{\log\log n}\} = \frac{1}{\sqrt{2\pi}} \int_{\omega_1}^{\omega_2} e^{-\frac{y^2}{2}}\,\mathrm{d}y, \forall \omega_1, \omega_2 \in \mathbb{R}. \tag{19}$$

Since all the proofs of this result are quite lengthy and involves **Sieve theory** which is a number-theoretic method used to give estimates on exressions of prime numbers/integers by eliminating unnecessary residue classes and thus it is out of scope of this talk. In fact, we will only give a sketch of the proof which involves some probabilistic methods. At this point it should be noted that this theorem is not a corollary of the **CLT**, although technically from **Kac's** original heuristic it follows that $\nu(n)$ is a sum of independent random variables with mean and variance $\log\log n$.

Proof. (Sketch) Let $K_n(\omega)$ denote the number of integers m, $1 \leq m \leq n$, for which

$$\nu(m) < \log\log n + \omega\sqrt{\log\log n},$$

and set $\sigma_n(\omega) = \frac{K_n(\omega)}{n}$. Then one can observe that $\sigma_n(\omega)$ is a distribution function. Next, denote $f_m = \frac{1}{\sqrt{n\log\log n}}(\nu(m) - \log\log n)$ and define $\beta(\omega) = \sum_{m=1}^{n} \delta_0(\omega - f_m)$ where $\delta_0(x)$ is the **Dirac-delta** function. Then we know that $\int g(\omega)\delta_0(\omega-\omega_0)\,\mathrm{d}\omega = g(\omega_0)$. Let $g(\omega) = \omega^2, \forall \omega \in \mathbb{R}$ and put $\omega_0 = f_m$, to obtain

$$\frac{1}{n\log\log n}\sum_{m=1}^{n}(\nu(m) - \log\log n)^2 = \int_{-\infty}^{\infty} \omega^2\,\mathrm{d}\sigma_n(\omega) \tag{20}$$

as $\frac{d}{d\omega}\sigma_n(\omega) = \frac{\beta(\omega)}{n}$. If we use the precise estimate

$$\sum_{p\leq n}\frac{1}{p} = \log\log n + C + \varepsilon_n, \quad \varepsilon_n \xrightarrow{n\to\infty} 0, \tag{21}$$

then from (14) and (15), it follows that

$$\lim_{n\to\infty}\int_{-\infty}^{\infty} \omega^2\,\mathrm{d}\sigma_n(\omega) = 1 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}}\,\mathrm{d}y. \tag{22}$$

Also, it is easy to observe that with the estimate given in (21), it follows that

$$\lim_{n\to\infty} \frac{1}{n\sqrt{\log\log n}} \sum_{m=1}^{n} (\nu(m) - \log\log n) = 0,$$

and hence

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \omega \, \mathrm{d}\sigma_n(\omega) = 0 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-\frac{y^2}{2}} \, \mathrm{d}y. \tag{23}$$

If we could prove that for every integer $k > 2$,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \omega^k \, \mathrm{d}\sigma_n(\omega) = 1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^k e^{-\frac{y^2}{2}} \, \mathrm{d}y, \tag{24}$$

it would follow that

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} e^{i\xi\omega} \, \sigma_{\mathrm{n}}(\omega) = e^{\frac{\xi^2}{2}}, \forall \xi \in \mathbb{R}$$

and hence that

$$\lim_{n\to\infty} \sigma_n(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\omega} e^{-\frac{y^2}{2}} \, \mathrm{d}y.$$

□

One can observe that (24) is equivalent to,

$$\frac{1}{n(\log\log n)^{\frac{k}{2}}} \sum_{m=1}^{n} (\nu(m) - \log\log n)^k = \int_{-\infty}^{\infty} y^k e^{-\frac{y^2}{2}} \, \mathrm{d}y. \tag{25}$$

But the proof of (25), depends on the asymptotics of $\sum_{p_{l_1}\cdots p_{l_k} < n} \frac{1}{p_{l_1}\cdots p_{l_k}}$ (recall as in **Turán's** proof in section 1 depended on an estimate of $\sum_{pq \leq n} \frac{1}{pq}$.) which could be achieved by **Sieve theory** and was proved by **Erdős**. **Turán, Alfred Rényi** and later **Heini Halberstam** (individually) generalized it for any additive number-theoretic function.

References

[1] Kac, Mark (1959). *Statistical Independence in Probability, Analysis and Number Theory*, John Wiley and Sons, Inc.

[2] Kac, Mark (1949.) *Probability methods in some problems of analysis and number theory*, Bull. Amer. Math. Soc. 55 (1949), 641–665.

[3] Kar, Arpita; Murty, M. Ram (2020) *The central limit theorem in algebra and number theory. Modular forms and related topics in number theory*, 101–124, Springer Proc. Math. Stat., 340, Springer, Singapore, [2020]